Ensemble methods for handwritten digit recognition

Hansen, Lars Kai; Liisberg, Christian; Salamon, P.

Published in:
Proceedings of the IEEE-SP Workshop Neural Networks for Signal Processing

Link to article, DOI:
10.1109/NNSP.1992.253679

Publication date:
1992

Document Version
Publisher's PDF, also known as Version of record

ENSEMBLE METHODS FOR HANDWRITTEN DIGIT RECOGNITION

L. K. Hansen[1], C. Liisberg[2], and P. Salamon[3]

Abstract. Neural network ensembles are applied to handwritten digit recognition. The invidual networks of the ensemble are combinations of sparse Look-Up Tables with random receptive fields. It is shown that the consensus of a group of networks outperform the best invidual of the ensemble and further we show that it is possible to estimate the ensemble performance as well as the learning curve, on a medium size database. In addition we present preliminary analysis of experiments on a large data base and show that *state of the art* performance can be obtained using the ensemble approach by optimizing the receptive fields.

INTRODUCTION

Recognition of handwritten digits is a serious, current candidate for a "real world" benchmark problem to assess pattern recognition methods: For a recent review see [4]. It has been the object of a recent state of the art application of neural networks [5].

Neural network ensembles were introduced recently as a means for improving network training and performance. The consensus of a neural network ensemble may outperform individual networks [1] and ensembles can be used to implement *oracle* functions [2]. Furthermore the consensus may be used for realization of fault tolerant neural network architectures [3]. Within the present system for recognition of handwritten digits, we find that the ensemble consensus outperform the best individual of the ensemble by $20-25\%$. However, due to correlation among errors made by the participating networks, the marginal benefit obtained by increasing the ensemble is low once the ensemble size is $\gtrsim 15$. Our findings are in line with the results obtained in [1]. We illustrate the theoretical tools for predicting the performance of the ensemble consensus, and we demonstrate the use of the ensemble as an *oracle* in its capacity of predicting the *learning curve*, ie. the number of test errors as a function of the number of training examples. This and other ensemble oracle functions were introduced by Salamon *et al.* [2]. Real world applications face the problem of

[1]CONNECT, Electronics Institute B349, The Technical University of Denmark, DK-2800 Lyngby Denmark, lars@eiffel.ei.dth.dk

[2]CONNECT, Dept. of Optics and Fluid Dynamics, Risø National Laboratory, DK-4000 Roskilde, liisberg@risoe.dk

[3]Dept. of Mathematical Sciences, San Diego State University, San Diego CA 92182 USA, salamon@math.sdsu.edu

0-7803-0557-4/92$03.00 © 1992

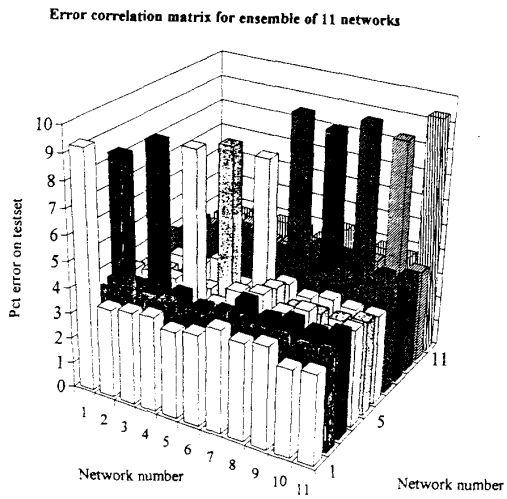

Figure 1: Correlation of test errors among members of an ensemble of 11 networks trained on a database of 3471 handwritten digits. The test set of 3500 examples was written by an independent group of 140 people. Note that the level in the diagonal is the average performance. If the errors were independent, the level outside the diagonal would be the square of the level in the diagonal.

noise (eg. mis-classifications or sloppy handwritings) and an important problem relates to the estimation of such noise levels. In this context we introduce, and use the ensemble to estimate, the *modelling deficiency*. This quantity is defined as the residual test error using the model and it is estimated as the frequency of consensus errors in an infinite ensemble. The modelling deficiency is determined by the model design as well as by the inherent noise level. As a specific result we present evidence that there is, on the average, only one dominant alternative (mis-classification) for each digit.

The present study is based on a pattern recognition strategy designed by Liisberg *et al.* [7]. The individual network device is a collection of *Look-Up Tables* (LUT's) with sparse random receptive fields. The randomness of the individual networks tends to differentiate their generalization (test) errors, creating an ideal setting for ensembles. By combining such LUT's, one may obtain *improved performance* by using the consensus of the ensemble. Figure 1 illustrates the level of correlation between networks for the particular application. The heights of the columns are the fractions of test examples with coincident errors in an ensemble of 11 networks.

In the next section we discuss the LUT-approach and section three reviews the basic notions of ensemble theory. In section four we discuss experiments on a 7000-digit database, while section five contains some concluding remarks as well as an outline of future work on extended databases. In particular we present a preliminary analysis of experiments on the NIST data base [8], we show that by using optimized receptive fields it is possible to obtain *state of*

the art performance.

SPARSE RANDOM LOOK-UP TABLES

The network system is a feed-forward net with one "hidden layer". The units of the hidden layer are equipped with sparse random receptive fields. Each hidden unit receives binary input from n_R input units. The activity pattern in the receptive field is interpreted as the bit-pattern of an address in a Look-Up Table. Each LUT has 2^{n_R} rows of M entries, with M being the number of output categories.

Training Phase. In the training phase an example, consisting of a binary image and the corresponding classification, is loaded on the network by incrementing the activity of the entry corresponding to the given particular address and the given output category. After a single pass through the training set the entries will have different activities reflecting the correlations among bit-patterns/addresses and classes.

The size of the receptive fields n_R determines together with M the number of entries in the LUT's. To ensure proper generalization we fix n_R by the heuristic: $n_R = \lfloor \log_2(m) - 1 \rfloor$ where m is the number of training examples, this leaves two examples pr. row on the average. A network consists of N_{LUT} Look-Up Tables. N_{LUT} is in turn determined by the constraint that all pixels should contribute: $N_{LUT} = \lceil n_{pixels}/n_R \rceil$ where n_{pixels} is the number of pixels in the visual field. The assignment of the receptive fields is a random permutation of the pixel indices, ensuring that all pixels participate in the classification.

Application of the Trained Network. In the application phase an example, ie. a binary image with unknown classification, produces a set specific bit-pattern in the receptive fields of the LUT's. Each pattern is translated into the proper address and the classification is given by the class of the particular address having the highest activity. The output of the LUT is fed to the output-layer where the activity of the given class is incremented by one. As a result of the LUT processing, each digit obtains a score. The network output is the digit obtaining the highest LUT activity.

ENSEMBLE METHODS

The consensus decision of an ensemble of pattern recognition devices may be more reliable than that of the individuals if two basic qualitative criteria are met: 1) The individual networks are performing reasonably well. 2) The errors of the different networks are to some degree independent. The necessity of the first criterion is obvious; there is no such thing as a free lunch. Provided that the second criterion also holds, the consensus decision is usually superior to *the best individual* even when there is a large range in individual performance [1].

Estimating Ensemble Performance. Correlation of errors among networks may be present for two reasons. There may be induced correlation in the way the networks err by the method used to create the networks. Secondly we have to face the fact that some examples of the pattern recognition problem at hand, are more difficult than others. The distribution of pattern difficulty is naturally discussed in terms of the ensemble. The *difficulty* θ of an example is defined as the fraction in a large ensemble of trained networks that misclassifies it. This induces a *difficulty distribution* $\mu(\theta)$ that gives the probability of seeing an example with difficulty θ in a random choice from the set of possible patterns which are candidates for classification. Using $\mu(\theta)$, we may estimate the performance of the plurality decision (where the option with the largest number of votes wins) using the tools proposed in [1].

In this presentation we consider plurality decisions among $M = 10$ categories and we will estimate the consensus performance taking the difficulty distribution into account. The prediction is based on the observation that once we have formulated the problem in the *difficulty representation*, the errors can be treated as if they were independent. We first compute the performance of an ensemble of N devices, on the slice of example space that has difficulty θ:

$$P^{N,M}_{plurality}(\theta) = \sum_{n_1=0}^{\lfloor 1+\frac{N-1}{M} \rfloor} \binom{N}{n_1} \theta^{n_1}(1-\theta)^{N-n_1} + \sum_{n_1=\lceil 1+\frac{N-1}{M} \rceil}^{N} \binom{N}{n_1} \theta^{n_1}(1-\theta)^{N-n_1} H(N,M,n_1) \quad (1)$$

with:

$$H(N,M,n_1) = 1 - \frac{\sum_{k=0}^{M-1}(-1)^k \binom{M-1}{k} \binom{N-n_1(k+1)+M-2}{N-n_1(k+1)}}{\binom{N-n_1+M-2}{M-2}} \quad (2)$$

and with the proviso that binomial coefficients with negative entries are zero. $P^{N,M}_{plurality}(\theta)$ is the probability that any one of the $M-1$ equally likely alternatives gets more votes than the correct alternative [1]. The formulas are best used with M as an adjustable parameter: the *effective degree of confusion*, M_{eff}. The value of M_{eff} is typically less than the actual number of output categories for the problem. Note that for $M = 2$, $H(N,2,n_1) = 0$ for all $n_1 > N/2$ and thus the second term on the RHS of equation (2) vanishes. To get the average performance we integrate the above expression using the difficulty distribution:

$$P^{N,M_{eff}} = \int P^{N,M_{eff}}_{plurality}(\theta)\mu(\theta)d\theta \tag{3}$$

Prediction of the Learning Curve. Applications tend to run short of examples. The necessary number of training examples is therefore a most important issue. A central result, obtained by Schwartz *et al.* [9], predicts the test-error as a function of the training set size. Schwartz *et al.* consider the distribution of generalization proficiencies g (test performances) of the ensemble of possible networks specified by (say) a given architecture. Assuming the distribution of generalization abilities, $\rho_{m_0}(g)$, to be given at some specific training set size m_0, the predicted distribution of networks compatible with a larger set of $m + m_0$ examples is approximately given by: $\rho_{m+m_0}(g) = g^m \rho_{m_0}(g) / \int_0^1 g^m \rho_{m_0}(g) dg$.

This in turn makes it possible to predict the learning curve from the given measured ρ_{m_0}:

$$E_{m+m_0} = 1 - \int_0^1 \rho_{m+m_0}(g)dg = 1 - \frac{\int_0^1 g^{m+1}\rho_{m_0}(g)dg}{\int_0^1 g^m \rho_{m_0}(g)dg} \tag{4}$$

To estimate ρ_{m_0} from a finite sample of networks, we apply a *Maximum Entropy* argument. We choose the distribution ρ_{m_0} with the largest entropy consistent with the values of three measurements: the mean ($\overline{g}$), the variance (v_g) and the range of generalizations ($[0, g_{max}]$). While $\overline{g}$ and v_g are readily computed from a given ensemble, we need to develop an estimate of g_{max}. We propose here to use the extrapolated performance of an infinite ensemble trained on the given training set by taking the limit $N \to \infty$ in equation (3). The resulting maximum entropy distribution is of the form: $\rho_{m_0}(g) = z^{-1} \exp\left(-(g-a)^2/2b\right)$ where the parameters a, b are determined by the constraint that the mean and variance are given by the measured values: $\overline{g}, v_g$, and where $z^{-1} \equiv z^{-1}(a, b, g_{max})$ ensures the proper normalization on the interval $[0, g_{max}]$. The resulting distribution may then be inserted in (4) to predict the learning curve.

EXPERIMENTAL

The primary body of experiments for this study is based on a digit database consisting of 6973 handwritten digits written by 280 people, who all had filled a one page form with preprinted boxes and rectangles for the digits. The digits were scanned as binary images with a resolution of 200 dots pr. inch, segmented and scaled to fit within a 16 by 16 grid. Digits whose width was bigger than their height were scaled so width equalled height, else the proportions were kept and the resulting digits left-adjusted. There is approximately the same number of examples of each of the digits $(0-9)$ in the database. The database has been inspected and cleaned for false classifications, segmentation errors and *dirt*. All digits are readable by a human, however for some of the digits the

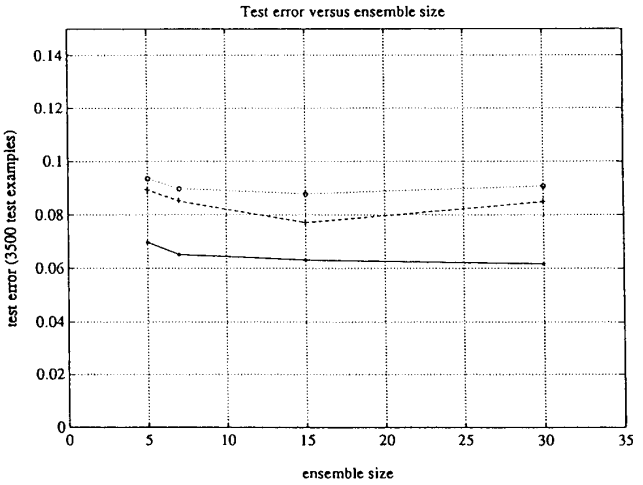

Figure 2: Test error of ensembles trained on 3471 handwritten digit. The average error (dotted), the ***best of ensemble*** (dashed), and the plurality consensus error rate (full line).

human readers needed to inspect the ***writing style context*** of other digits. The test set and training set were written by two different groups of 140 people each. Two series of experiments were conducted, in the first series network ensembles with $5, 7, 15$, and 30 members were trained on a test set of 3471 digits. In the second experiment a 7-member ensemble was trained on training sets of size: 200, 500, 1000, 1500, 2000, 2500, 3000, 3471. In both cases the test set was 3500 digits. Consensus decisions were implemented using the plurality scheme where the digit that collects the maximum number of ensemble votes wins. A vote is forced from each network, i.e., there is no rejection at the individual member or at the consensus level.

We crossvalidate the networks on the 3500-digit test set. The ensemble average, the ***best of ensemble*** and the ***plurality performance*** are depicted in fig. 2 as a function of the ensemble size. Our first observation is that the consensus outperform the best individual of the ensemble by $20 - 25\%$. The benefit of increasing the ensemble size beyond 15 members is marginal, however. We use eqs. (1)-(3) to estimate the consensus performance of larger ensembles based on measurements from the 7-member ensemble. The difficulty distribution was recorded and used for prediction. In fig. 3 we show the predictions using various effective degrees of confusion M_{eff}. The conclusion is that the best fit to the performance of the 7-member ensemble is obtained by $M_{eff} = 2$. ***This indicates that, when a network makes an error, there is only one dominant alternative available for each example.*** This conclusion is corroborated by a direct inspection of the distribution of errors. In particular we counted the number of wrong alternatives that had obtained more votes than the correct classification. This count estimated the average number of alternatives considered by the networks to be 1.4.

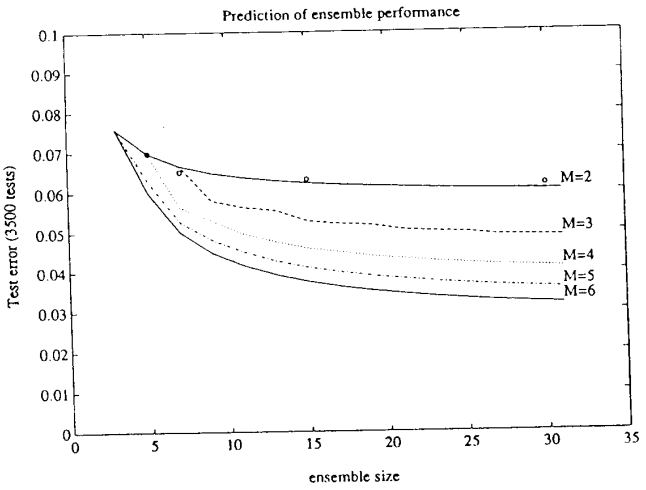

Figure 3: Theoretical prediction of consensus performance for varying *effective degrees of confusion*, M. Experimental data (o).

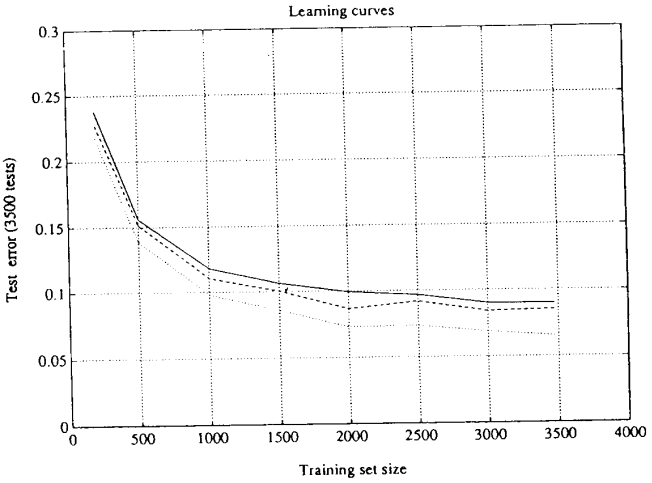

Figure 4: Learning curves for the average of the ensemble (full line), the best of ensemble (dashed line), and for the consensus (dotted line).

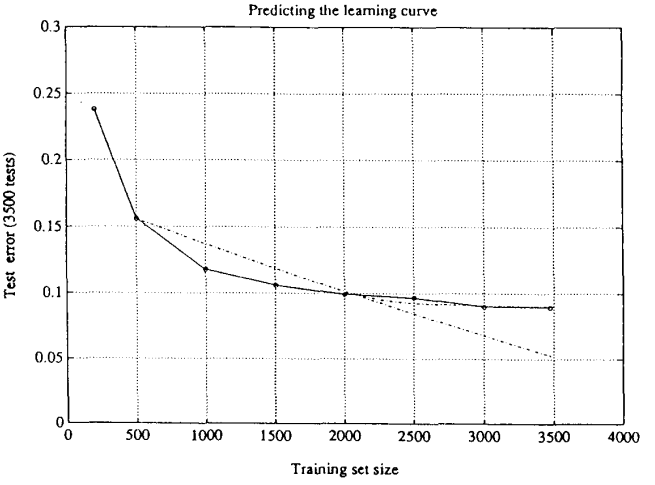

Figure 5: Theoretical prediction of the learning curve based on the data of the 7-member ensemble trained on 500 examples (dash-dotted lines). The lower prediction in the plot is based on $g_{max} = 1.0$, while the upper curve is based on the estimate of the modelling deficiency: $g_{max} = 0.91$.

The learning curves for the average and for the plurality consensus of the 7-member ensemble are presented in fig. 4. We use the ensemble to compute the mean and variance based on 500 training examples. The estimated maximum generalization ability of the model, given the noise level, is derived from the extrapolated performance of an infinite ensemble trained on 500 examples. As noted above, using $M_{eff} = 2$, makes the second term on the RHS of equation (2) vanish. Taking the limit of the resulting $P^{N,M_{eff}}_{plurality}(\theta)$ as $N \rightarrow \infty$ gives $P^{N,M_{eff}}_{plurality}(\theta) = 1$ if $\theta > 0.5$ and $P^{N,M_{eff}}_{plurality}(\theta) = 0$ if $\theta < 0.5$. Using this in equation (3) gives the simple expression[1]: $P^{\infty,2} = \int_{\theta=0.5}^{1.0} \mu(\theta) d\theta$ Based on this estimator we find the *modelling deficiency* to be: $1 - g_{max} \sim P^{\infty,2} = 0.09$. In fig. 5 we show the learning curve and the two predictions based on $g_{max} = 0.91$ and on $g_{max} = 1.0$ respectively. The extrapolated error rate is remarkably close to the experimental limit of the average performance if we use the estimated modelling deficiency.

In order to improve the classification performance of the individual networks we invoke a modified design based on optimized LUT's. In this case the networks are constructed in an iterative scheme. While constructing the networks we simulate a *leave one out* cross validation procedure of a pool of candidate LUT's. The result of the cross validation test can be computed in a very compact form using a score table quantifying the network activity for each example [10]. The successfull candidate, which together with the current network leads to the optimal validation result, is included in the network, and the procedure is carried on until the gain is negligible. Furthermore we apply a *ensemble reject criterion* simply enforcing that the plurality winner beats the *runner up* by a

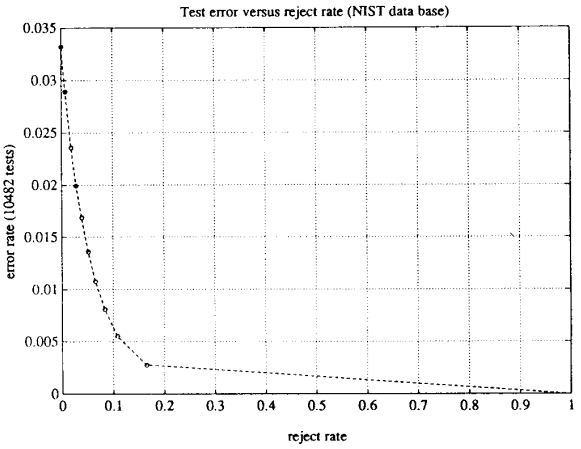

Figure 6: Experiment on the NIST data base: The error rate of an 18 member ensemble, versus fraction of rejected patterns.

given margin. We trained ensembles with up to 30 networks of 30 LUT's each having 15 unit receptive fields, on digits of the NIST database [8]. To facilitate training we divided the training set among the ensemble members using 7000 examples pr. member. In fig. 6. we present the error rate versus fraction of rejected patterns of an 18 member ensemble on a test set of 10482 examples. While it is impossible to compare directly with other reported results since these are based on different data bases, it is evident that the above results are at level with state of the art results as reported by [5, 6].

CONCLUSION

In this work we have discussed the use of ensemble methods for the identification of handwritten digits. The problem is of great practical importance. We have shown that it is possible to improve performance significantly by introducing moderate-size ensembles; in particular we found a $20-25\%$ improvement. Our ensemble random LUT's, when trained on a medium size database, reaches a performance (without rejects) of 94% correct classification on digits written by a independent group of people. As a comparison, the state of the art system developed in [5] obtained an error rate of 1% with 9% rejects. In preliminary analysis of ensembles of optimized LUT networks trained on the large NIST data base we reach an error rate of 0.8% with 9% rejects. The notion of modelling deficiency has been introduced and an estimator for it has been proposed. Using the estimated modelling deficiency we are able to predict the learning curve. We have presented arguments that the networks tend to confuse *pairs* of classifications. This, however, is not simply explained by a pair-wise confusion of digits. Rather some subset of the instances of a given digit is confused with subsets of the instances of another digit. This merits a more detailed future

study, using the available international digit databases.

ACKNOWLEDGEMENTS

We thank Nils Hoffmann and Jan Larsen for helpfull comments on the manuscript. This work is supported by the Danish Natural Science and Technical Research Counsils through the Computational Neural Network Center (CONNECT).

References

[1] L.K Hansen and P. Salamon: *Neural Network Ensembles*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **12**, 993-1001 (1990).

[2] P. Salamon, L. K. Hansen, B. E. Felts III., and C. Svarer: *The Ensemble Oracle*, AMSE Conference on Neural Networks. San Diego 1991.

[3] L. K. Hansen and P. Salamon: *Self-Repair in Neural Network Ensembles*, AMSE Conference on Neural Networks. San Diego 1991.

[4] V.K. Govindan and A.P. Shivaprasad: *Character recognition - a review*, Pattern Recognition **23** (1990).

[5] Y. Le Cun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, and L.D. Jakel: *Handwritten Digit Recognition with a Back-Propagation Network*, In Advances in Neural Information Processing Systems II (Denver 1989) ed. D.S.Touretzsky, 396-404. San Mateo: Morgan Kaufman. (1990)

[6] Y. Lee: *Handwritten Digit Recognition Using K Nearest-Neighbor, Radial-Basis Function,and Backpropagation Neural Networks* Neural Computation **3**, 440-449, (1991)

[7] C. Liisberg: *Low-priced and robust expert systems are possible using neural networks and minimal entropy coding*, To appear in: Expert systems with applications. 1991 Pergamon Press.

[8] National Institute of Standards and Technology: *NIST Special Data Base 3, Handwritten Segmented Characters of Binary Images*, HWSC Rel. **4-1.1** (1992).

[9] D.B. Schwartz, V.K. Salalam, S.A. Solla, and J.S. Denker: *Exhaustive Learning*, Neural Computation **2**, 371-382 (1990).

[10] Thomas Martini Jørgensen and Christian Liisberg: Private Communication.